DEVELOPMENT OF EFFICIENT MULTI-LEVEL DISCRETE WAVELET TRANSFORM HARDWARE ARCHITECTURE FOR IMAGE COMPRESSION

by

KHAMEES KHALAF HASAN AL-JUMAIL

Thesis submitted in fulfillment of the
requirements for the degree of
Doctor of Philosophy

July 2015

ACKNOWLEDGEMENTS

First and foremost, I would like to thank and praise Allah, the almighty Merciful, who bestowed upon me by facilitating the successful accomplishment of this PhD research work. Respectfully, I acknowledge that my success is not only related to me but its possession to all the people whom guide me, teach me, pray for me and encourage me. From them, I learned the knowledge, patient, wisdom, humility and also how to gain access to my goals.

I would like to seize this opportunity to specially acknowledge to my supervisor, Assoc. Prof. Dr. Umi Kalthum bt Ngah for her invaluable suggestions, dedication support, constructive effort and beneficial comments that have remarkably influenced to bring this work into light. Her advices and words of encouragement helped me to overcome most of the difficulties I have faced. I am proud to conduct this research under her supervision.

The acknowledgement would be incomplete if I do not express my sincere gratitude to my co-supervisors, Assoc. Prof. Dr. Mohd Fadzli Mohd Salleh, for his support and encouragement to discuss and share his views about any issues related to this research. He always finds the time for listening to the little problems and roadblocks that unavoidably crop up in the course of performing research.

Special thanks to Universiti Sains Malaysia for their cooperation and providing the facilities required for this research. Also, special unreserved appreciation goes to the friendly staff of the school of Electrical and Electronic Engineering, for their cooperation and friendly attitude. My sincere appreciation is extended to Dr. Satam H. Hasan for his invaluable support and encouragement during the study. I would like to extend my deepest appreciation to my best friends, the Phd students Mr. Luqman Mahmood and Mr. Salam Kareem for their support, prayers and endless friendship and encouragement.

I would like to express the deepest appreciation to the Ministry of High Education and Scientific Research in Iraq for its support. I wish acknowledge to University of Tikrit, Faculty of Engineering and Department of Electrical and Electronic Engineering for their cooperation and facilities on approval for giving me the opportunity to seek for this Phd degree.

Last but not the least, I would like to dedicate my gratitude to the dearest and nearest people to my heart, my family for their unremitting support, encouragement and boundless patience with me throughout the years of the research, especially my beloved wife and my wonderful kids, who patiently underwent the alienation and distance from the family without complaints and are supportive of me. This thesis is for you. Finally, to all my family members, I am forever indebted for your understanding, support, endless patience and encouragement when it matters the most. You are all like bright candles in the dark days and difficult times, so thank again.

TABLE OF CONTENTS

page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

1-D	One Dimension
2-D	Two Dimensions
AI	Algebraic Integer
ASICs	Application Specific Integrated Circuits
BB	Block-Based
CAD	Computer Aided Design
CB	Connection Blocks
CL	Computational Load
CMOS	Complementary Metal Oxide Semiconductor
CODEC	Coding and Decoding
CP	Column Processor
CODEC	Coding and Decoding
DCT	Discrete Cosine Transform
DE	Development and Education
DSPs	Digital Signal Processors
DWT	Discrete Wavelet Transform
EBCOT	Embedded Block Coding with Optimized Truncation
EZW	Embedded Zerotree Wavelet
FB	Filter Bank
FDWT	Forward Discrete Wavelet Transform
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FSM	Finite State Machine
FPGAs	Field-programmable gate arrays

FU	Filtering Unit
GPPs	General Purpose Processors
HDL	Hardware Description Language
HUE	Hardware Utilization Efficiency
HVS	Human Visual System
HWT	Haar Wavelet Transform
IDWT	Inverse Discrete Wavelet Transform
IRSA	Interlaced Read Scan Algorithm
JPEG	Joint Photographic Experts Group
LABs	Logic Array Blocks
LB	Line-Based
LBs	Logic Blocks
LDWT	Lifting based Discrete Wavelet Transform
LEs	Logic Elements
LS	Lifting Scheme
LUTs	Lookup Tables
MA	Multiplies and Adds
MAC	Multiply and Accumulate
MFA	Modified Flipping Architecture
MFHWT	Modified Fast Haar Wavelet Transform
MBW-PKCM	Modified Baugh-Wooley Pipelined Constant Coefficient Multiplier
MPEG	Moving Picture Experts Group
MRA	Multi Resolution Analysis
MSE	Mean Square Error
NSFU	Non-Separable Filtering Unit

PA	Pyramidal Algorithm
PAR	Place and Route
PE's	Processing Elements
PLDs	Programmable logic devices
PLLs	Phase Locked Loops
PSNR	Peak Signal to Noise Ratio
RC	Row Column
RPA	Recursive Pyramidal Algorithm
RTL	Register Transfer Level
SB	Switch Block
SD	Secure Digital
SDRAM	Synchronous Dynamic Random Access Memory
SFU	Separable Filtering Unit
SPIHT	Set Partitioning in Hierarchical Trees
SRAM	Static Random Access Memory
STFT	Short-Time Fourier Transform
TS1D	Type State one dimensional
TSMC	Taiwan Semiconductor Manufacturing Company
UP	University Program
VHDL	(VHSIC) Hardware Description Language
VHSIC	Very High Speed Integrated Circuit
VGA	Video Graphics Array
WTU	Wavelet Transform Unit

LIST OF PUBLICATIONS

International Journals

HASAN, K. K., NGAH, U. K. & SALLEH, M. F. M. 2014. Efficient Hardware-Based Image Compression Schemes for Wireless Sensor Networks: A Survey. *Wireless personal communications,* vol.77**,** pp. 1415-1436.

International Proceedings:

HASAN, K. K., NGAH, U. K. & SALLEH, M. F. M. 2013a, Multilevel decomposition Discrete Wavelet Transform for hardware image compression architectures applications, *IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, 29 Nov. - 1 Dec. 2013, Penang, Malaysia, pp.315-320,

HASAN, K. K., NGAH, U. K. & SALLEH, M. F. M. 2013b, Low complexity image compression architecture based on lifting wavelet transform and embedded hierarchical structures, *IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, 29 Nov. - 1 Dec. 2013, Penang, Malaysia, pp. 305-309.

HASAN, K. K., NGAH, U. K. & SALLEH, M. F. M. 2012, The most proper wavelet filters in low-complexity and an embedded hierarchical image compression structures for wireless sensor network implementation requirements, *IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, 23-25 Nov. 2012, Penang, Malaysia, pp. 137-142.
.

Symposium Papers:

HASAN, K. K., NGAH, U. K. & SALLEH, M. F. M. 2013, Energy-Efficient and Low Complexity Image Compression Scheme Based on Lifting Wavelet Transform and Embedded Hierarchical Structures for Wireless Devices Applications, *the 4 th Symposium on Image Processing, Image Analysis and RealTime Imaging (IPIARTI), and the 1 st Symposium on Acoustic, Speech and Signal Processing (SASSP), IEEE Signal Processing Society Malaysia Chapter* , 09 May 2013, Universiti Tenaga Nasional, Putrajaya Campus.

HASAN, K. K., NGAH, U. K. & SALLEH, M. F. M. 2011, Discrete Cosine Transform Based Image Compression in Visual Sensor Networks, *the 3 rd Postgraduate Colloquium of the School of Electrical & Electronics Postgraduate Colloquium (EEPC2011)*, 2-4 Dec. 2011, Dusun Eco Resort, Bentong, Pahang, Malaysia.

PEMBANGUNAN CEKAP BAGI SENI BINA PERKAKASAN JELMAAN GELOMBANG KECIL DISKRET BERBILANG ARAS UNTUK PEMAMPATAN IMEJ

ABSTRAK

Berfokuskan pengkomputeran intensif dalam gelombang kecil diskret (DWT), reka bentuk seni bina perkakasan efisen bagi pengkomputeran laju menjadi imperatif terutamanya dalam aplikasi masa nyata. Keseluruhan objektif menunjukkan pemetaan tugas pengkomputeran berkaitan dengan pelbagai tahap resolusi 5/3 DWT merupakan model matematik pertama dan kemudiannya reka bentuk tahap DWT dikenal pasti melalui pendapat tentang pengelakan "subband" frekuensi tinggi pengkomputeran. Tambahan pula, asas jelmaan gelombang kecil Haar (HWT) digunakan untuk perbandingan. FWDT yang dicadangkan dan reka bentuk perkakasan penapis IDWT memberi keputusan hampir sama berbanding dengan model MATLAB bagi tujuh aras penguraian DWT. Simulasi dilakukan menggunakan imej skala kelabu untuk saiz yang berbeza bagi mengesahkan reka bentuk cadangan dan mencapai prestasi kelajuan sesuai dengan jumlah aplikasi masa nyata. Empat versi reka bentuk dihasilkan dalam kajian ini iaitu saiz imej piksel 64x64, 128×128, 256×256, dan 512× 512. Litar FDWT yang dicadangkan daripada seni bina 5/3 2-D DWT boleh memproses imej 256 × 256 dalam 2.07301ms, yang empat kali lebih cepat daripada JPEG2000 dan pelaksanaan HWT FPGA dengan kurang penggunaan perkakasan. Penapis 5/3 FDWT yang dicadangkan menghasilkan 127 keping perkakasan logik dan kawasan elemen berdaftar yang mengandungi kurang daripada 1% papan pembangunan Altera DE2 kawasan perkakasan Cyclone II FPGA.

DEVELOPMENT OF EFFICIENT MULTI-LEVEL DISCRETE WAVELET TRANSFORM HARDWARE ARCHITECTURE FOR IMAGE COMPRESSION

ABSTRACT

Focusing on the intensive computations involved in the discrete wavelet transform (DWT), the design of efficient hardware architectures for a fast computation of the transform has become imperative, especially for real-time applications. With this overall objective, the mapping of the computational tasks associated with the various resolution levels of the 5/3 DWT is first mathematically modeled, and then a modified computation of the DWT stages are explored from the standpoint of evading computing high frequency subbands. Furthermore, a Haar wavelet transform (HWT) is used for comparison. The proposed forward DWT (FDWT) and its inverse DWT (IDWT) hardware architecture filter generated similar results compared to the MATLAB model for the seven levels of DWT decomposition. Simulations were performed using grayscale images of different sizes to validate the proposed design and attain speed performance appropriate for a number of real-time applications. Four versions of the design are developed in this study, 64×64, 128×128, 256×256, and 512× 512 pixels image sizes. FDWT circuit of the proposed 5/3 2-D DWT architecture can process a 256×256 image in 2.07301 ms, which is at least four times faster than that of the other JPEG2000 and HWT FPGA implementation with less hardware utilization. The proposed 5/3 FDWT filter produced 127 slices of hardware logic and register element area, which comprises less than 1% of the Altera DE2 development board Cyclone II FPGA hardware area.

CHAPTER 1

INTRODUCTION

1.1 Background

The objective of the image compression procedure is to reduce the redundancy or irrelevant aspects of the image data in order to save or transmit data in an ordered form (Wu and Abouzeid, 2005). A model of a typical image compression system consists of three major components; removal or decrease in source data redundancy with the objective of reducing the duplicate information, reduction in entropy so as to get rid of irrelevant information not recognized by the human visual system (HVS), and finally, installing an effective entropy encoding as outlined in Figure 1.1 (Wu and Abouzeid, 2005).

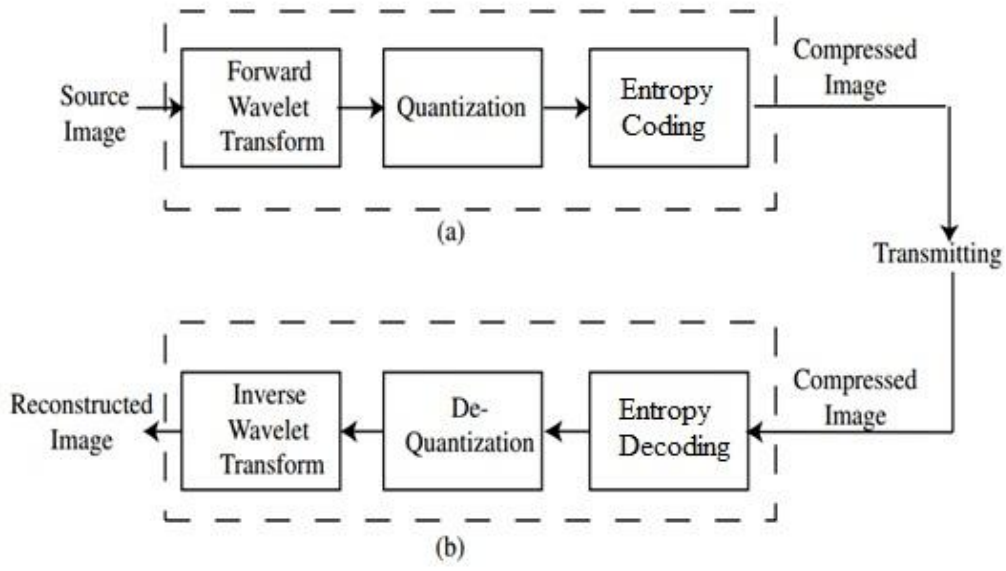

Figure 1.1: Functional wavelet-based image compression (a) encoder (b) decoder (Wu and Abouzeid, 2005)

Data redundancy consists of three source types, and they include, spatial, spectral and temporal redundancies (Acharya and Tsai, 2005). In still images, spatial redundancy is the most prevalent, given the self-similarity within the non-edge smooth area in the image (Acharya and Tsai, 2005). Spectral redundancy combines the various spectral bands or color planes. Temporal correlation has been used at length in video compression of successive frames (Mamun et al., 2014). At present, transform-based coding methods have garnered lots of interest due to its ability to decorrelate image data by removing spatial and spectral redundancies in still images (Mandal, 2003). The widely recognized transform-based techniques used in the redundancy reduction procedure include; Fourier-related transforms such as discrete cosine transform (DCT) and the multi-resolution Discrete Wavelet Transform (DWT). Theoretically, these techniques have the capacity to develop more compact representations of the pixel information of the original data set (Mandal, 2003).

The reduction in entropy involves lowering the entropy of the transformed data significantly so as to allocate smaller bits for transmission or storage through quantization process. Redundancy embedded in an image surpassing the human sensorial frequency threshold will be wholly imperceptible to the human senses and is indiscernible by the eyes in standard visual processing (Gonzalez et al., 2002). Hence, it is not obvious and can be removed without visually impinging on the picture quality. Quantization is a component of the supplementary preprocessing block in the image compression framework. This step can be omitted if information loss on media could cause superfluous results. Thus, the most appropriate compression technique should be applied according to the application and the end

user of the compressed file (Salomon, 2004). A compression is referred to as a lossless technique if the reconstructed data is an exact replication with features consistent with the original data. On the contrary, it is a lossy technique when the reconstructed data is not exactly similar to original data (Salomon, 2004). Afterward, the quantization procedure is accompanied by lossless encoding with the use of several entropy methods to efficiently characterize the quantized data for storage or transmission (Mandal, 2003).

The decompression system is simply an inverse procedure. Firstly, the compressed image is decoded to generate the quantized coefficients. The inverse quantization step is subsequently applied on these quantized coefficients to generate an approximate of the transformed coefficients. The quantized transformed coefficients are then inversely modified to develop an approximate adaptation of the original data (Acharya and Tsai, 2005; Salomon, 2004). This study focuses on the DWT process, which is the computation-intensive primary component of the wavelet-based image compression system that supports power consumption. The study asserts that optimal algorithmic elements incorporated into the wavelet transform step can viably enhance the overall effectiveness and speed performance requisites of the entire compression system.

1.2 Discrete Wavelet Transform Fundamentals and Computations

In recent times, the Discrete Wavelet Transform (DWT) has been widely used in several fields such as image compression (Boix and Canto, 2010), speech analysis (Jayakumar and Babu Anto, 2013) and pattern recognition (Pogrebnyak et al., 2014)

because of its ability to decompose a signal at multiple resolution levels with adaptive time-frequency window. The DWT decomposes a signal into components of different octaves or frequency bands by selecting suitable scaling and shifting factors where the small and large scaling factors correspond to the fine and coarse details of the signal, respectively, while the shifting factor to the time or space localization of the signal (Vetterli and Kovačević, 1995). This transform differs from others, such as Fourier or cosine transforms where the signals are only represented in frequency domain, given the DWT decomposes a signal so that it is well represented and discretized in both time and frequency domains. Thus, the time information contained in the DWT does not dissipate in the transformed signal, which is a focal point in signal analysis, particularly for signals characterized with non stationary or transitory features (Sifuzzaman et al., 2009).

DWT is a multiple-level decomposition transform that can be implemented by repeating a process wherein a fully scalable window is shifted along the dimensions of the signal thereby shortening the window size with each repetition (Kaiser, 2010). The DWT computing process involves repeatedly executing a set of instructions developed in software programs. The software implementation for the computation of the DWT is flexible in setting different values of the parameters of the transform and varying the codes for the algorithms (Subbarayan and Karthick Ramanathan, 2009). However, despite the effort given to the design of software algorithms and optimized codes for their implementations, no general-purpose or digital signal processor can provide a performance in terms of the computing speed and resource optimization that can possibly be attained with hardware implementation (Chaver et al., 2002; Shahbahrami et al., 2008).In hardware

implementation, the computation of the DWT is performed with a custom hardware circuit, making it possible to deal with the requirements of precise applications such as the speed, power or size of the circuit.

Over the years, several design efforts have been made on the development of architectures for the DWT computation that focus on such requirements of applications (Weeks and Bayoumi, 2003; Benderli et al., 2003; Chang and Gaofeng, 2006). Nonetheless, the bulk of these applications involve large-volume data such as image or video. Consequently, it remains to be a challenging task to model high speed, low-power and area-efficient architectures to implement the DWT computation for real-time image compression applications. In the following sections, a definition of image compression is provided. The motivations for developing multiple resolution level DWT hardware architecture are given. Then, the research objectives and scope are explained, which is followed by the research methodology overview. Finally, an outline of the organization of this thesis is presented.

1.3 Problems Statement

Currently, a diverse number of architectures have been proposed designed to provide high-speed computation of the DWT using resource-efficient hardware. The problems to dealing with hardware utilization and computation time issues related to wavelet based image compression algorithms in real time environments are summarized as:

(a) The previous architectures (Lewis and Knowles, 1991; Grzeszczak et al., 1996; Denk and Parhi, 1998; Hung et al., 2001; Guo et al., 2001; Movva and Srinivasan, 2003; Liao et al., 2004; Uzun and Amira, 2004;

Nayak, 2005; Meher et al., 2008) utilize a single processor to execute the computations of all the resolution levels of the DWT, generally in accordance with the recursive pyramid algorithm (RPA) (Vishwanath, 1994). As expected, by using a single processor in these architectures, the computations of the different resolution levels of the DWT are executed sequentially, given that the computation at one resolution level requires the output data from its preceding level as a requisite parameter. Therefore, even though these architectures have simple designs and low hardware complexities, they do not provide fast DWT computation, making them unappealing for real-time applications.

To surmount the constraint of slow computation, architectures that utilize two or more parallel processors have been proposed (Parhi and Nishitani, 1993; Chakrabarti and Vishwanath, 1995; Chakrabarti and Mumford, 1996; Marino et al., 1999; Chen et al., 2001; Wu and Chen, 2001; Masud and McCanny, 2004; Farahani and Eshghi, 2006). In these architectures, the computation involved in one level can be performed by more than one processor thereby increasing the overall processing speed of the DWT computation. Although this type of architectures provides parallelism to the computations associated with a given resolution level, they do not have parallelism between the resolution levels. To further enhance the parallelism for the DWT computation, and hence the computational speed, architectures that incorporate several pipelined stages, each carrying out the task of one or more resolution levels of the DWT, have been proposed (Jer Min et al., 1999; Marino, 2000; Marino et al., 2000; Marino, 2001; Chen, 2004).

However, the design of these architectures which is incorporating a number of parallelisms in the computations associated with the DWT, do not provide high throughput and overall shorter duration of computation.

In addition to that, it is apparent that numerous pipeline architectures concentrating on providing high computational speed or efficient hardware utilization have been proposed in the literature do not offer speed which is not proportionate with the amount of hardware utilized by them to perform the filtering operations. Moreover, no systematic approach seems to exist in mapping of the DWT multiple resolution levels in order to reduce the computation time to minimal level and efficiently maximize hardware utilization. Such an approach will ensure the effective implementation of the computation for real-time wavelet-based image compression applications.

(b) The commonly applied image compression and decompression standard is the Joint Photographic Experts Group (JPEG), which utilizes discrete cosine transform (DCT) kernel. DCT-based algorithms are high-speed algorithms with minimal computation complexities and low storage capacity applicable to 8-by-8 blocks of raw image data. Nonetheless, JPEG has a number of constraints, mainly in low bit-rate applications. Spatial correlation from DCT adjoining blocks is disregarded, which frequently results in blocking artifacts and subsequently have an effect on the effectiveness of the coding system (Luo and Ward, 2003; Quijas and Fuentes, 2014). An interpreter may not correctly recognize the reconstructed image due to the degradation of image quality (Luo and

Ward, 2003). Therefore, it requiring eliminating all limitations and incorporating improved features. Multi Resolution Analysis (MRA) algorithms developed with the DWT resolves the limitations of the Fourier transform and its derivatives (Mallat, 1989a). Wavelets surpass other more conventional decomposition techniques such as the discrete Fourier transform DFT and DCT with basis functions more appropriate for representing images. The basis functions associated with wavelet decomposition usually have long support for representing slow variations in an image and short support for efficiently represent sharp transitions i.e., edges. The problem is, DFT and DCT basis functions have support over the entire image, making it hard to represent both slow variations and edges efficiently (Adams, 2013). Therefore another way needs to be explored. DWT has been confirmed to be exceptionally successful for transform-based image compression, where it substitutes the DCT in JPEG2000 and MPEG4 image and video compression standards (Adams, 2013).

Since the inception of JPEG2000 image coding standard in 2002, cost effectiveness and real-time constraints continue to be the major impediments to hardware realization of JPEG2000 standard into end user products (Mansouri et al., 2009). The DWT has been executed with the conventional convolution method and Lifting Scheme (LS) spatial approach. LS is a very flexible system, with simple adders and shifters substituting multipliers. Consequently, it basically decreases the amount of multiplication and accumulation entailed in analyzing a DWT using a convolution approach (Sweldens, 1996; Daubechies and Sweldens,

1998). The LS high algorithmic DWT performance in image compression validates its application as the essential part of the JPEG2000 standard. The Daubechies 9/7 and LeGall 5/3 LS DWT filters are used as the default JPEG2000 standard filters for lossy and lossless image compression, respectively (Taubman et al., 2002). The 9/7 transform incorporates floating-point coefficients in its transform filters, although it is a complex computational architecture for embedded parallel processing (Krishnaiah et al., 2012), while the typical 5/3 wavelet filter coefficients are approximated by dyadic rational numbers, where all division operations can be executed as bit shifts, thus ensuring a speedy computational method (Le Gall and Tabatabai, 1988; Adams and Kossentni, 2000). Such inherent hardware features that exist between the coefficients of the filter make 5/3 LS -based schemes attractive to further enhance the operating speed for applications were requiring real-time performances. The execution of complex algorithms is so computationally demanding that specific function hardware solutions needs to be developed (Benkrid et al., 2002). JPEG2000 standard uses the symmetric extension at the boundaries or margins to remove these edge effects (Christopoulos et al., 2000). Hence, extended computations or clock cycles at the start and at the end are dissipated in processing each row and column of the image (Po-Chih et al., 2002).

Moreover, no mathematical derivation approach seems to exist in mapping of the computational tasks and data access loads associated with the various resolution levels of the of the JPEG2000 standard DWT.

(c) Since, the amount of bandwidth resource is very limited, there is the need to design communication system for image transmission that save the transmission bandwidth and at the same time delivering a good quality of the received data. It is imperative to remove redundant DWT computations.

(d) Efficient fast algorithm (folded pyramidal computing scheme) for the computation of discrete wavelet coefficients makes a wavelet transform based encoder computationally efficient. Different filter banks with different characteristics can be used as comparisons to demonstrate the potential of the DWT hardware core to perform image processing techniques.

1.4 Research Objectives

The main aim of this research is to enhance the potential and efficacy of DWT computation-intensive algorithm implementation, particularly for efficient wavelet based image compression applications. The key objectives of this study can be generally summarized as follows:

i. To develop an enhanced 5/3 LS-based DWT multilevel decomposition hardware architecture;
ii. To develop the tools for evaluating the performance of the proposed scheme by deriving the mathematical expressions of the multilevel decomposition computational and data access loads;
iii. To develop an image transmission scheme based on the proposed 5/3 LS-

based DWT multilevel decomposition scheme that can significantly save the computation as well as communication energy that isneeded;

iv. To evaluate the usefulness of effective inclusive hardware acceleration of the proposed 5/3 LS as well as analyze and compare its effectiveness with HWT and the standard JPEG2000 Lee Gall 5/3 LS models.

A step-by-step approach is taken in this research to achieve the above objectives. The scope of the thesis and research methodology is explained in the next sections.

1.5 Scope of the Thesis

The DWT operation determines the volume of computations in successive resolution levels. This thesis embarks on designing a fast hardware compatible and resource-efficient embedded extension architecture for the computation of 1-D and 2-D DWTs. Based on this general objective, a comprehensive mathematical model of DWT power consumption is initially developed by computing the computational and data access loads associated with the multi-resolution levels of the 5/3 LS-based DWT decomposition process. Therefore, a detailed modification in the DWT decomposition process is implemented to avoid the computation of high frequency subbands. It is envisaged that the modified DWT structure provides a solution to the image quality limitations with significant reduction in power consumption. An overall view of DWT-based image compression chain is outside the scope of this study.

1.6 Overview of Methodology

A general overview of the research flow is depicted in this Section. The scope of this research is focused on the DWT multilevel decomposition computations, which is an essential part of the overall image compression system as explained in Section 1.2. Given the scope of this research, the DWT as LS framework is reviewed and analyzed as an alternative to the less-effective convolution-based methods. The 5/3 LS filter is selected as the main supporting module for developing the DWT multilevel decomposition, because of several salient features, particularly its ability to combat the symmetric boundary extension difficulty and the normal filter coefficients nature, which offers a fast computation option. In order to enhance the 5/3 LS filter performance, two approaches are taken: (i) enhancing the 5/3 LS symmetric extension that saves clock cycles and supplementary computations in a less complex way; and (ii) developing a simple 5/3 LS filter framework to evade the computation of the high-pass coefficients samples generated from high frequency subbands during the DWT multilevel decomposition process.

Apart from the low power embedded extension 5/3 LS filter, in-depth investigations are also conducted for the simple HWT developed using the linear algebra framework to resolve power consumption problems of the multilevel decomposition DWT in real time environments. In the high speed DWT multilevel decomposition stage, the challenge is to maximize the operational clock frequency and minimize the number of clock cycles necessary for the 2-D DWT computation. In addition, the hardware resources are taken into consideration. As a result, a number of enhancements to the DWT multilevel decomposition stage are proposed.

1.7 Thesis Outline

This thesis is organized in accordance with the objectives of the study. Comprehensive review of discrete mathematical models for the computation of 1-D and 2-D wavelet transforms along with the methods of computation is described in chapter 2. The review encompasses the existing architectures for the computation of these transforms. Chapter 3 comprises a concise appraisal of theoretical approaches and mathematical formulations for developing a scheme to design resource efficient hardware architectures for fast computation of the 2-D DWT using low power embedded extension 5/3 LS filter. The second part of this chapter focuses on deriving computations to eliminate DWT high frequency subbands so as to lower power consumption. A comprehensive study, for the design and FPGA implementation of HWT architecture to illustrate and validate the proposed scheme, as well as make comparisons with other existing schemes for the design of architectures for the 2-D DWT computation is presented in Chapter 4. To demonstrate the applicability of the low power 5/3 LS model devised in Chapter 3, and the HWT proposed in Chapter 4, the discussion on experimental results obtained are presented in Chapter 5, as an attempt to ascertain the efficacy of the low power 5/3 LS model in real-world environment. The conclusions of the study are outlined in Chapter 6. The contributions of this research as well as a number of areas proposed for future pursuit are also presented in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

As pointed out in chapter 1, the central focus of this study is to examine the computation efficiency of multilevel LS-based DWT. In accordance with this theme, this chapter provides background information required for the architectural development of the 1-D and 2-D DWT carried out in the subsequent chapters. Firstly, the mathematical formulations of 1-D and 2-D DWT are outlined, along with the methods of their computations. This is followed by a review of the different presently applied architectures, grouped into single-processor, parallel-processor and pipeline architectures for the 1-D and 2-D DWT computations. Based on the review of related studies, few algorithms have been proposed for computation of multilevel DWT decomposition in MRA. Multilevel 2-D DWT computation can be executed using PA, Recursive Pyramid Algorithm (RPA) and folded scheme. These algorithms are concisely analyzed, with a summary of the review presented at the end of this chapter.

2.2 Fundamentals of the DWT

The term ‘wavelet’ was first used in the early 1980s by the French researchers, Jean Morlet and Alex Grossman, who used the French word ‘ondelette’ meaning a “small wave”. Sometime afterward, the word was translated into English, from “onde” to “wave”, subsequently acquiring the name wavelet (Vetterli and Kovačević, 1995). As the name indicates, wavelet is a waveform that displays an oscillating wavelike attribute that tends to be asymmetric, irregular and limited in

duration with an average value of zero (Vetterli and Kovačević, 1995). The wavelet transform (WT) was developed to resolve the constraint of short time Fourier transform (STFT) and non-stationary signals (Gabor, 1946). While STFT presents an invariable resolution at all frequencies, the WT uses MRA through which diverse frequencies are examined with varying resolutions. The wavelet analysis is performed similarly to the STFT analysis. The signal to be analyzed is multiplied with a wavelet function in the same way it is multiplied with a window function in STFT, and the transform is subsequently computed for every segment (Adams, 2013). The time information contained in WT is derived by shifting the wavelet over the signal, while the frequencies are modified through the processes of contraction and dilatation of the wavelet function. The continuous wavelet transform (CWT) recovers the time-frequency content information with an enhanced resolution in contrast with the STFT (Louis et al. 1997). The redundancy of information and the huge complex computations needed to compute all likely translations and scales of CWT limits its application. A substitute to this analysis is the discretization of the scale and translation factors using DWT. There are numerous approaches to initiating the notion of DWT, the major ones are the decomposition bands and the decomposition pyramid, which were first introduced in the late 70's (Rioul and Vetterli, 1991).

DWT was originally developed in 1976 to decompose discrete time signals (Croisier, 1976). In the same year, a similar technique for coding speech signals called subband coding was created as well (Crochiere et al., 1976). Later on in 1983, the pyramidal coding analysis, a technique analogous to subband coding was created (Burt and Adelson, 1983). DWT was also introduced by Mallat in his multi-

resolution theory (Mallat, 1989a). DWT is a mathematical technique that offers a novel method for signal processing and the decomposition of a discrete signal in the time domain using shifted and translated versions of a single basis function, referred to as prototype wavelet (Mallat, 1989a). Following the development of the Mallat multi-resolution theory, several changes were made to the subband coding scheme in 1989, which incorporated the removal of existing redundant peripheral features in the pyramidal coding scheme (Vetterli and Le Gall, 1989). The theory and implementation of DWT were further analyzed and broadened in contemporary studies (Mallat, 1989b; Daubachies, 1992; Meyer, 1993; Vetterli and Kovačević, 1995).

DWT provides broad range of more valuable features when compared to other transforms like discrete Fourier transforms (DFT), discrete cosine transform (DCT) and discrete sine transform (DST). These features include; adaptive time-frequency windows, lower aliasing distortion for signal processing applications and inherent scalability (Grzesczak et al., 1996). These features facilitate the implementation of one dimensional (1-D) and two dimensional (2-D) DWTs in different applications, which include numerical analysis (Beylkin et al., 1992), statistics (Stoksik et al., 1994;), pattern recognition (Kronland-Martinet et al., 1987; Pittner and Kamarthi, 1999), image coding (Sodagar et al., 1999; Taubman, 2000), signal analysis (Akansu and Haddad, 2000) and biomedicine (Senhadji et al., 1994; Bagheri Hamaneh et al., 2014). A number of algorithms and computation schemes have been proposed in the last 30 years in order to improve the efficiency of 1-D DWT and 2-D DWT hardware implementation. The various computation schemes normally applied in hardware implementation are concisely discussed in the next section.

2.3 Computation Scheme for 1-D DWT

In Mallat algorithm, the traditional DWT can be efficiently implemented using sub band coding scheme. DWT decomposes the input signal into two subbands referred to as low-pass subband and high-pass subband (Mallat, 1989a). The low-pass and high-pass subband components of a given DWT decomposition level are derived by filtering the input signal using a pair of low pass filter (LPF) and high-pass filter (HPF). The low-pass and high-pass filter pair forms a quadrature mirror filter (QMF) structure to attain perfect signal reconstruction (Gonzalez and Woods, 2002). The low-pass and high-pass filters are also defined as finite impulse response (FIR) filter.

As illustrated in Figure 2.1, the low-pass filter output stream is sub-sampled (down-sampled) to obtain the low-pass subband output $y_l(\mathrm{n})$. Likewise, the highpass filter output stream is decimated by a factor of two to gain the high-pass subband output $y_h(\mathrm{n})$ by basically deleting the alternate output samples in each stream(Mallat, 1989a). The first level 1-D DWT decomposition can be illustrated with the filtering unit (FU) block diagram shown in Figure 2.1.

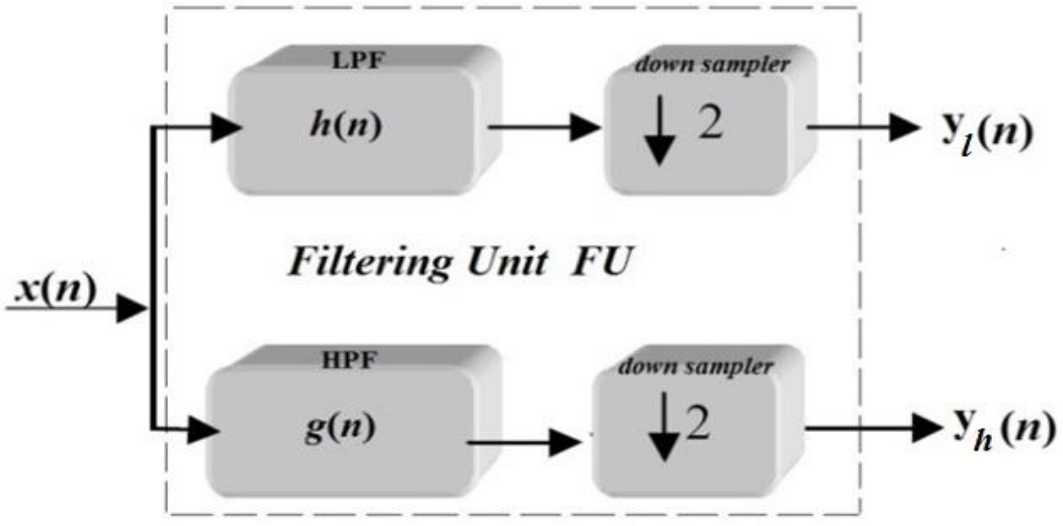

Figure 2.1**:** Computation of one level 1-D DWT

The low-pass and high-pass filter outputs are determined by means of two computation schemes: convolution scheme and lifting scheme. These computation schemes are explained in sections 2.3.1 and 2.3.2.

2.3.1 Convolution Scheme

In conventional convolution schemes applied in the computation of 1-D-DWT, the low-pass and high-pass filter output of an FU are calculated using Mallat's algorithm (Mallat, 1989a). The 1-D DWT computation is equal to the function of two channel down-sampled FIR filter computation. The FU of 1-D DWT comprises a pair of filters LPF and HPF, and a pair of down samplers. Mathematically, the 1-D DWT filtering operation of a signal entails the convolution of the signal with the impulse response of the LPF and HPF, and then down sample operation. To avoid computing terms not needed, these two steps of convolution and down-sampling should be combined as (Jensen and la Cour-Harbo, 2001):

$$y_l[\mathrm{n}] = \sum_{\mathrm{i}=0}^{\mathrm{k}_1-1} h[i]x[2n-i] \tag{2. 1}$$

$$y_h[\mathrm{n}] = \sum_{\mathrm{i}=0}^{\mathrm{k}_2-1} g[i]x[2n-i] \tag{2.2}$$

where, k_1 denotes the length of low-pass filter, k_2 signifies the length of high-pass filter, $x(n)$ represents the input signal, $y_l(\mathrm{n})$ and $y_h(\mathrm{n})$ symbolize the low-pass and high-pass subband components, and $h(n)$ and $g(n)$ mean the analysis low-pass and high-pass filter coefficients of wavelet, respectively. The original signal can be reconstructed using synthesis FU.

In (Mallat, 1989a) algorithm, the components required for synthesis FU include interpolator and filters. The low-pass and high-pass subband coefficients, $y_l(\mathrm{n})$ and $y_h(\mathrm{n})$ are applied inversely in the synthesis stage by up-sampling and filtering with low pass and high pass synthesis FU. Filtered outputs are subsequently added to obtain the final output. The perfect reconstruction condition ensures the absence of distortion and aliasing in the reconstructed data, which have been removed by the filters (Gonzalez and Woods, 2002; Adams, 2013).

Wavelet filters are grouped into, orthogonal and bi-orthogonal wavelets. The wavelet filter coefficients that satisfy the orthogonal property are referred to as orthogonal wavelet. The orthogonal low-pass and high-pass filters are, asymmetric with similar lengths, while the low-pass and the high-pass filters of biorthogonal wavelet are symmetric with contrasting lengths (Rao and Bopardikar, 1997). Biorthogonal wavelets produce invertible matrices and perfect reconstruction (Strang and Nguyen, 1996). In addition, they are valuable to image processing because of their symmetrical coefficients. JPEG2000 is composed of two types of biorthogonal wavelet filter banks: Cohen Daubechies Feauveau (CDF) 9/7 biorthogonal filter and LeGall 5/3 biorthogonal filter for lossy and lossless image compression respectively (Marcellin, 2002).

The distinctive LeGall 5/3 biorthogonal wavelet filter encompasses rational coefficients, which offers a rapid computational approach to signal decomposition (Le Gall and Tabatabai, 1988). The $h(n)$ 5-tap LPF coefficients of 5/3 filters include [-1/8 2/8 6/8 2/8 -1/8] while the $g(n)$ 3-tap HPF coefficients of 5/3 filters comprise [-1/2 1 -1/2]. Both $h(n)$ and $g(n)$ analysis filters are symmetric parameters. Given that

all filter coefficients are computed using dyadic numbers, every division operations can be executed as bit shifts to enhance computational speed (Adams and Kossentni, 2000). For the synthesis filter pair for inverse transformation, the low-pass FIR filter comprises 3 filter coefficients are [1/2 1 1/2] while the high-pass FIR filter contains 5 coefficients are [-1/8 -2/8 6/8 -2/8 -1/8]. The corresponding synthesis filters were generated according to optimum settings for reconstruction using complementary filters (Acharya and Chakrabarti, 2006).

The fundamental 9/7 Daubechies architecture comprising FIR filter coefficients for a 9-tap low-pass filter and a 7-tap high-pass filter is shown in Figure 2.2. DWT filtering is usually the convolution operation, that is, FIR filtering which can be implemented with 14 adders, 16 multipliers and 9 registers structure (Silva and Bampi, 2005). However, this architecture requires high area costs for the parallel implementation. Therefore, a number of algorithms were initiated to reduce this area cost, for instance the lifting algorithm (Silva and Bampi, 2005).

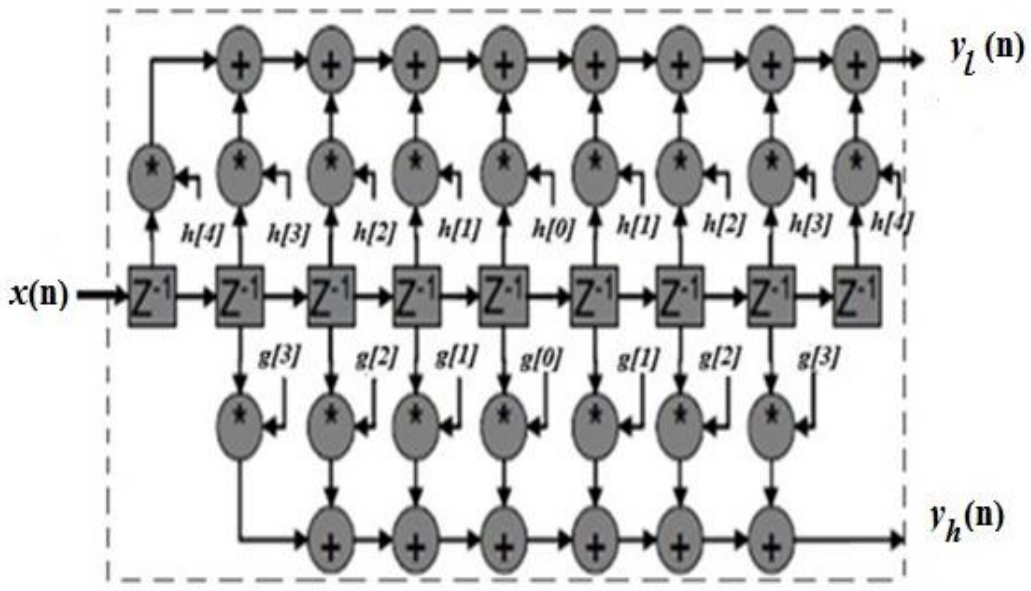

Figure 2.2: DWT by 9/7 tap Daubechies FIR filter (Silva and Bampi, 2005)

Generally, in the LS -based technique, the amount of arithmetic operations is asymptotically half of the convolution FU number using a spatial domain analysis that can be executed in low memory systems. Accordingly, lifting has been recommended for execution of DWT in JPEG2000 standard (Mansouri et al., 2009), as detailed in the next section.

2.3.2 Lifting Scheme (LS)

The LS was first put forward by Sweldens, (1996). Based on the lifting scheme, computation of any FU of 1-D DWT can be incorporated into lifting steps. The fundamental principle of lifting scheme is to factorize the polyphase matrix $\hat{P}(z)$ of wavelet filters into a series of alternating upper and lower triangular matrices and a constant diagonal matrix (Sweldens, 1996). This results in wavelet computation through banded-matrix multiplications (Daubechies and Sweldens, 1998). The lifting based DWT has several valuable intrinsic features such as symmetric forward and inverse transform, in-place computation, and integer-to-integer transform and demands a reduced volume of computation as compared to convolution based DWT (Acharya and Chakrabarti, 2006).

To comprehend the primary step in lifting theorem of splitting a specific signal into it's even and odd components mathematically, it is essential to determine Z-domain representation of the polyphase components. H(z) and G(z) are considered the system function of low-pass and high-pass wavelet filters. H(z) can be divided discretely into He(z) and Ho(z), where He(z) and Ho(z) symbolize the system functions of even and odd part of the impulse response h(n) of low-pass wavelet

filter. Equally, Ge(z) and Go(z) correspond to the system functions of even and odd part of the impulse response g(n) of high-pass wavelet filter. The system function of FU can be denoted by polyphase matrix analysis as (Mansouri et al., 2009):

$$\hat{P}(z) = \begin{bmatrix} He(z) & Ho(z) \\ Ge(z) & Go(z) \end{bmatrix} \tag{2.3}$$

The polyphase matrix $\hat{P}(z)$ can be factorized into lower, upper matrix as (Daubechies and Sweldens, 1998):

$$\hat{P}(z) = \left\{ \prod_{i=1}^{m} \begin{bmatrix} 1 & Si(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ Ti(z) & 1 \end{bmatrix} \right\} \begin{bmatrix} K & 0 \\ 0 & 1/K \end{bmatrix} \tag{2.4}$$

where *K* is the scaling constant, *Si*(*z*) and *Ti*(*z*) are the system function of predict and update unit of *i*-th lifting step (for $1 \leq i \leq m$) as shown in Figure 2.3 (Acharya and Chakrabarti, 2006).

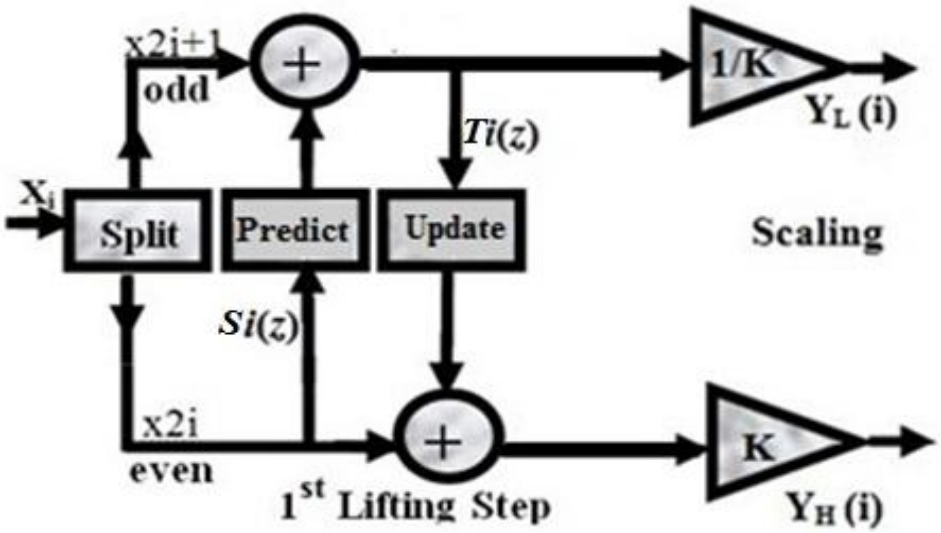

Figure 2.3: LS 1-D-DWT block diagram (Acharya and Chakrabarti, 2006)

Each predict and update stage represents one lifting step of DWT. For example: lifting computation of an FU using 9/7 biorthogonal wavelet filter is outlined in four lifting step as (Daubechies and Sweldens, 1998; Rein and Reisslein, 2011):

$$\hat{P}(z) = \begin{bmatrix} 1 & \alpha(1+z^{-1}) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \beta(1+z) & 1 \end{bmatrix} \begin{bmatrix} 1 & \gamma(1+z^{-1}) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \delta(1+z) & 1 \end{bmatrix} \begin{bmatrix} \mathrm{K} & 0 \\ 0 & 1/\mathrm{K} \end{bmatrix} \quad (2.5)$$

where α= −1.5861343420693648, β= −0.0529801185718856, γ= 0.8829110755411875, δ= 0.4435068520511142 and K= 1.1496043988602418 are lifting parameters constants which can be derived through a factoring algorithm computation of a 9/7 biorthogonal wavelet coefficient FU (Daubechies and Sweldens, 1998; Rein and Reisslein, 2011). Similarly, lifting computation of a conventional LeGall 5/3 biorthogonal FU which leads to the polyphase matrix representation of the 5/3 analysis filter bank is given as (Andra et al., 2002; Acharya and Chakrabarti, 2006):

$$\hat{P}(z) = \begin{bmatrix} he(z) & ho(z) \\ ge(z) & go(z) \end{bmatrix} = \begin{bmatrix} -\frac{1}{8}z^{-1} + \frac{3}{4} - \frac{1}{8}z & \frac{1}{4} + \frac{1}{4}z \\ -\frac{1}{2}z^{-1} - \frac{1}{2} & 1 \end{bmatrix} \quad (2.6)$$

In line with the above equations, the decomposition of the analysis filter bank into a series of two upper and lower triangular matrix multiplication is specified as (Daubechies and Sweldens, 1998):

$$\hat{P}(z) = \begin{bmatrix} 1 & \frac{1}{4}(1+z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{1}{2}(1+z^{-1}) & 1 \end{bmatrix} \quad (2.7)$$

The process of computing the upper triangular matrix is referred to as the primal lifting step or update $T_i(z)$ step and is considered in related literature as the process of lifting the low-pass subband even terms with the support of the high-pass subband odd terms (Andra et al., 2002; Acharya and Chakrabarti, 2006). Similarly,

computation of the lower triangular matrix is referred to as dual lifting step or predict $S_i(z)$ step which involves lifting the high-pass subband odd terms with the assistance of the even terms low-pass subband (Andra et al., 2002; Acharya and Chakrabarti, 2006). Therefore, for the LS of the 5/3 filter-bank comprising only one predict-and-update step, the two basic multiplications (by (α= -1/2) and by (β=1/4)) lifting constants, are implemented as predict and update lifting steps coefficients (Daubechies and Sweldens, 1998). For the 5/3 LS filter, the division by two or by four operations can be basically performed with shifting one bit or two bits in the predict step or the update respectively (Adams and Kossentni, 2000). The operation of the 9/7 filter is a more complex process. In the contrast, the 5/3 filter does not need a scaling step, where the constant K is equal to unity as presented in the Figure 2.4 (Dia et al., 2009).

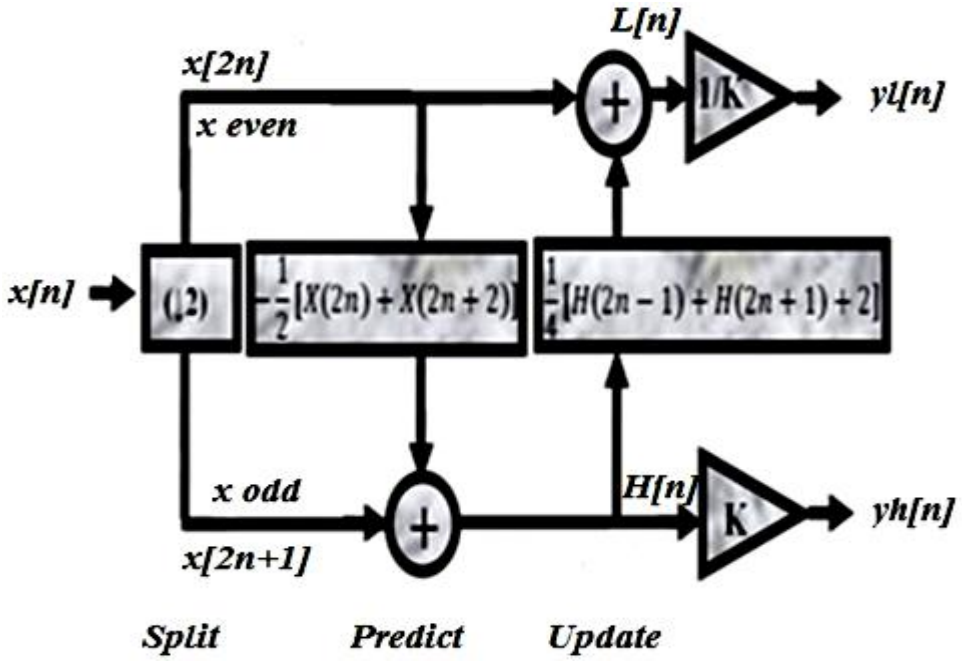

Figure 2.4: 1-D-DWT LS decomposition of 5/3 filter block diagram (Dia et al., 2009)